Look Who is Talking: Soundbite Speaker Name Recognition in Broadcast News Speech

Feifan Liu, Yang Liu
Department of Computer Science
The University of Texas at Dallas, Richardson, TX
{ffliu,yangl}@hlt.utdallas.edu

Abstract

Speaker name recognition plays an important role in many spoken language applications, such as rich transcription, information extraction, question answering, and opinion mining. In this paper, we developed an SVM-based classification framework to determine the speaker names for those included speech segments in broadcast news speech, called soundbites. We evaluated a variety of features with different feature selection strategies. Experiments on Mandarin broadcast news speech show that using our proposed approach, the soundbite speaker name recognition (SSNR) accuracy is 68.9% on our blind test set, an absolute 10% improvement compared to a baseline system, which chooses the person name closest to the soundbite.

1 Introduction

Broadcast news (BN) speech often contains speech or interview quotations from specific speakers other than reporters and anchors in a show. Identifying speaker names for these speech segmentations, called soundbites (Maskey and Hirschberg, 2006), is useful for many speech processing applications, e.g., question answering, opinion mining for a specific person. This has recently received increasing attention in programs such as the DARPA GALE program, where one query template is about a person's opinion or statement.

Previous work in this line includes speaker role detection (e.g., Liu, 2006; Maskey and Hirschberg, 2006) and speaker diarization (e.g., Canseco et al., 2005). In this paper, we formulate the problem of SSNR as a traditional classification task, and proposed an SVM-based identification framework to explore rich linguistic features. Experiments on Mandarin BN speech have shown that our proposed approach significantly outperforms the baseline system, which chooses the closest name as the speaker for a soundbite.

2 Related Work

To our knowledge, no research has yet been conducted on soundbite speaker name identification in Mandarin BN domain. However, this work is related to some extent to speaker role identification, speaker diarization, and named entity recognition.

Speaker role identification attempts to classify speech segments based on the speakers' role (anchor, reporter, or others). Barzilay et al. (2000) used BoosTexter and the maximum entropy model for this task in English BN corpus, obtaining a classification accuracy of about 80% compared to the chance of 35%. Liu (2006) combined a generative HMM approach with the conditional maximum entropy method to detect speaker roles in Mandarin BN, reporting a classification accuracy of 81.97% against the baseline of around 50%. In Maskey and Hirschberg (2006), the task is to recognize soundbites (which make up of a large portion of the "other" role category in Liu (2006)). They achieved a recognition accuracy of 67.4% in the English BN domain. Different from their work, our goal is to identify the person who spoke those soundbites, i.e., associate each soundbite with a speaker name if any.

Speaker diarization in BN aims to find speaker changes, group the same speakers together, and recognize speaker names. It is an important component for rich transcription (e.g., in the DARPA EARS program). So far most work in this area has only focused on speaker segmentation and clustering, and not included name recognition. However, Canseco et al. (2005) were able to successfully use linguistic information (e.g., related to person names) to improve performance of BN speaker segmentation and clustering.

This work is also related to named entity recognition (NER), especially person names. There has been a large amount of research efforts on NER; however, instead of

Proceedings of NAACL HLT 2007, Companion Volume, pages 101–104,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics

recognizing all the names in a document, our task is to find the speaker for a particular speech segment.

3 Framework for Soundbite Speaker Name Recognition (SSNR)

Figure 1 shows our system diagram. SSNR is conducted using the speech transcripts, assuming the soundbite segments are provided. After running NER in the transcripts, we obtain candidate person names. For a soundbite, we use the name hypotheses from the region both before and after the soundbite. A 'region' is defined based on the turn and topic segmentation information. To determine which name among the candidates is the corresponding speaker for the soundbite, we recast this problem as a binary classification problem for every candidate name and the soundbite, which we call an instance. A positive tag for an instance means that the name is the soundbite speaker. Each instance has an associated feature vector, described further in the following section. Note that if a name occurs more than once, only one instance is created for it.

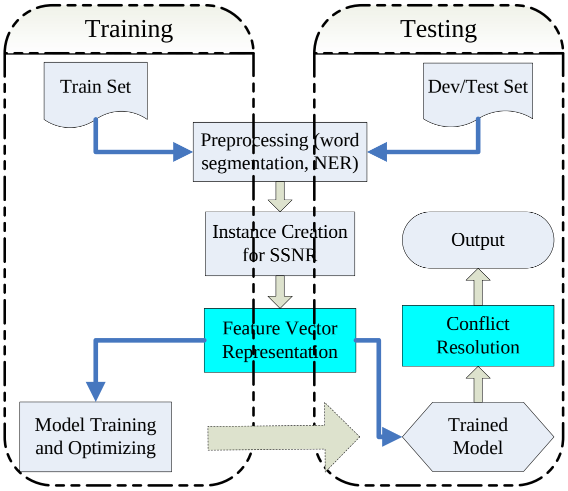
Training
Testing
Train Set
Dev/Test Set
Preprocessing (word segmentation, NER)
Instance Creation for SSNR
Output
Feature Vector Representation
Conflict Resolution
Model Training and Optimizing
Trained Model

Figure 1. System diagram for SSNR.

Any classification approach can be used in this general framework for SSNR. We choose to use an SVM classifier in our experiments because of its superior performance in many classification tasks.

3.1 Features

The features that we have explored can be grouped into three categories.

Positional Features (PF)

- PF-1: the position of the candidate name relative to the soundbite. We hypothesize that names closer to a soundbite are more likely to be the soundbite speaker. This feature value can be 'last', 'first', 'mid', or 'unique'. For example, 'last' for a candidate before a soundbite means that it is the closest name among the hypotheses before the soundbite. 'Unique' indicates that the candidate is the only person name in the region before or after the soundbite. Note that if a candidate name occurs more than once, the PF-1 feature corresponds to the closest name to the soundbite.
- PF-2: the position of a name in its sentence. Typically a name appearing earlier in a sentence (e.g., a subject) is more likely to be quoted later.
- PF-3: an indicator feature to show where the name has occurred, before, inside, or after the soundbite. We added this because it is rare that a name inside a soundbite is the speaker of that soundbite.
- PF-4: an indicator to denote if a candidate is in the last sentence just before the soundbite turn, or is in the first sentence just after the soundbite turn.

Frequency Features (Freq)

We hypothesize that a name with more occurrences might be an important subject and thus more likely to be the speaker of the soundbite, therefore we include the frequency of a candidate name in the feature set.

Lexical Features (LF)

In order to capture the cue words around the soundbite speaker names in the transcripts, we included unigram features. For example, "pre_word+1=说/said" denotes that the candidate name is followed by the word '说/said', and that 'pre' means this happens in the region before the soundbite.

3.2 Conflict Resolution

Another component in the system diagram that is worth pointing out is 'conflict resolution'. Since our approach treats each candidate name as a separate classification task, we need to post-process the cases where there are multiple or no positive hypotheses for a soundbite during testing. To resolve this situation, we choose the instance with the best confidence value from the classifier.

4 Experiments

4.1 Experimental Setup

We use the TDT4 Mandarin broadcast news data in our experiment. The data set consists of about 170 hours (336 shows) of news speech from different sources. Speaker turns and soundbite segment information were annotated manually in the transcripts. Our current study only uses the soundbites that have a human-labeled speaker name in the surrounding transcripts. There are 1292 such soundbites in our corpus. We put aside 1/10 of the data as the development set, another 1/10 as the test set, and used the rest as our training set. All the transcripts were automatically tagged with named entities using the NYU tagger (Ji and Grishman, 2005). For the classifier, we used the libSVM toolkit (Chang and Lin, 2001) and the RBF kernel in our experiments.

A reasonable baseline for SSNR is to choose the closest person name before a soundbite as its speaker. We will compare our system performance to this baseline approach.

We used two performance metrics in our experiments. First is the instance classification accuracy (*CA*) for the candidate names in the framework of the binary classification task. Second, we compute name recognition accuracy (*RA*) for the soundbites as follows:

$$RA = \frac{\#\ of\ Soundbites\ with\ Correct\ Names}{\#\ of\ Soundbites\ \ in\ Files}$$

4.2 Effects of Different Manually Selected Feature Subsets

We used 10-fold cross validation on the training set to evaluate the effect of different features and also for parameter optimization. Table 1 shows the instance classification results. “PF, Freq, LF” are the features described in Section 3.1. “LF-before” means the unigram features before the soundbites. “All-before” denotes using all the features before the soundbites.

Feature Subsets	Optimized Para.		*CA* (%)
	C	G	
PF-1	0.125	2	83.48
+PF-2	2048	1.22e-4	85.62
+PF-3	2048	4.88e-4	85.79
+PF-4	2	0.5	86.18
+Freq	2	0.5	86.18
+LF-before	32	7.81e-3	88.44
+LF-after i.e., All features	8	0.0313	88.44
All-before	8	0.0313	88.03

Table 1. Instance classification accuracy (*CA)* using different feature sets. C and G are the optimized parameters in the SVM model.

We notice that the system performance generally improves with incrementally expended feature sets, yielding an accuracy of 88.44% using all the features. Some features seem not helpful to system performance, such as “Freq” and “LF-after”. Using all the features before the soundbites achieves comparable performance to using all the features, indicating that the region before a soundbite contributes more than that after it. This is expected since the reporters typically have already mentioned the person’s name before a soundbite. In addition, we evaluated some compound features using our current feature definition, but adding those did not improve the system performance.

4.3 Automatic Feature Selection

We also performed automatic feature selection for the SVM model based on the F-score criterion (Chen and Lin, 2006). There are 6048 features in total in our system. Figure 2 shows the classification performance in the training set using different number of features via automatic feature selection.

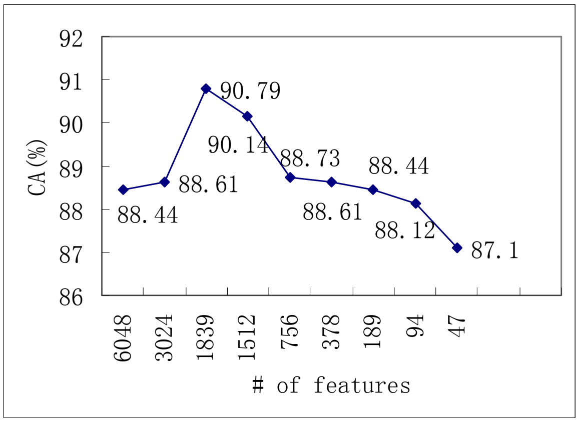

Figure 2. Instance classification accuracy (*CA*) using F-score based feature selection.

We can see that automatic feature selection further improves the classification performance (2.36% higher accuracy than that in Table 1). Table 2 lists some of the top features based on their F-scores. Consistent with our expectation, we observe that position related features, as well as cue words, are good indicators for SSNR.

Feature	F-score
Justbeforeturn (PF-4)	0.3543
pre_contextpos=last (PF-1)	0.2857
pre_senpos=unique (PF-2)	0.0631
pre_word+1=“上午/morning” (LF)	0.0475
pre_word+1= “说/said” (LF)	0.0399
bool_pre=1 (PF-3)	0.0353
Justafterturn (PF-4)	0.0349
pre_contextpos=mid (PF-1)	0.0329
post_contextpos=first (PF-1)	0.0323
pre_word+1= “今天/today” (LF)	0.0288
pre_word-1=“记者/reporter” (LF)	0.0251
pre_word+1=“表示/express” (LF)	0.0246

Table 2. Top features ordered by F-score values.

4.4 Performance on Development Set

Up to now our focus has been on feature selection based on instance classification accuracy. Since our ultimate goal is to identify soundbite speaker names, we chose several promising configurations based on the results above to apply to the development set and evaluate the soundbite name recognition accuracy. Results using the two metrics are presented in Table 3.

Feature Set	*CA* (%)	*RA* (%)
Baseline	84.0	59.3
PF	86.7	54.2
PF+Freq	86.7	60.4
PF+Freq+LF-before	87.8	63.5
PF+Freq+LF-before +LF-after (ALL)	88.3	67.7
Top 1512 by f-score	85.6	62.5
Top 1839 by f-score	85.4	60.4

Table 3. Results on the dev set using two metrics: instance classification accuracy (*CA*), and soundbite name recognition accuracy (*RA*). The oracle *RA* is 79.1%.

Table 3 shows that using all the features (ALL) performs the best, yielding an improvement of 4.3% and 8.4% compared to the baseline in term of the *CA* and *RA* respectively. However, using the automatically selected feature sets (the last two rows in Table 3) only slightly outperforms the baseline. This suggests that the F-score based feature selection strategy on the training set may not generalize well. Interestingly, "Freq" and "LF-after" features show some useful contribution (the 4th and 6th row in Table 3) respectively on the development set, different from the results on the training set using 10-fold cross validation. The results using the two metrics also show that they are not always correlated.

Because of the possible NER errors, we also measure the oracle *RA*, defined as the percent of the soundbites for which the correct speaker name (based on NER) appears in the region surrounding the soundbite. The oracle *RA* on this data set is 79.1%. We also notice that 8.3% of the soundbites do not have the correct name hypothesis due to an NER boundary error, and that 12.5% is because of missing errors.

We used the method as described in Section 3.2 to resolve conflicts for the results shown in Table 3. In addition, we evaluated another approach—we resort to the baseline (i.e., chose the name that is closest to the soundbite) for those soundbites that have multiple or no positive hypothesis. Our experiments on the development set showed this approach degrades system performance (e.g., *RA* of around 61% using all the features).

4.5 Results on Blind Test Set

Finally, we applied the all-feature configuration to our blind test data and obtained the results as shown in Table 4. Using all the features significantly outperforms the baseline. The gain is slightly better than that on the development set, although the oracle accuracy is also higher on the test set.

	CA (%)	*RA* (oracle: 85.8%)
Baseline	81.3	58.4
All feature	85.1	68.9

Table 4. Results on the test set.

5 Conclusion

We proposed an SVM-based approach for soundbite speaker name recognition and examined various linguistic features. Experiments in Mandarin BN corpus show that our approach yields an identification accuracy of 68.9%, significantly better than 58.4% from the baseline.

Our future work will focus on exploring more useful features, such as part-of-speech and semantic features. In addition, we plan to test this framework using automatic speech recognition output, speaker segmentation, and soundbite segment detection.

6 Acknowledgement

We thank Sameer Maskey, Julia Hirschberg, and Mari Ostendorf for useful discussions, and Heng Ji for sharing the Mandarin named entity tagger. This work is supported by DARPA under Contract No. HR0011-06-C-0023. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

S. Maskey and J. Hirschberg. 2006. Soundbite Detection in Broadcast News Domain. In *Proc. of INTERSPEECH2006*. pp: 1543-1546.

Y. Liu. 2006. Initial Study on Automatic Identification of Speaker Role in Broadcast News Speech. In *Proc. of HLT-NAACL*. pp: 81-84.

R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. 2000. The Rules Behind Roles: Identifying Speaker Role in Radio Broadcasts. In *Proc. of AAAI*.

L. Canseco, L. Lamel, and J.-L. Gauvain. 2005. A Comparative Study Using Manual and Automatic Transcriptions for Diarization. In *Proc. of ASRU*.

H. Ji and R. Grishman. 2005. Improving Name Tagging by Reference Resolution and Relation Detection. In *Proc. of ACL*. pp: 411-418.

Y.-W. Chen and C.-J. Lin. 2006. Combining SVMs with Various Feature Selection Strategies. *Feature Extraction, Foundations and Applications*, Springer.

C. Chang and C. Lin. 2001. LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.